文章编号：1000-8608(2015)02-0223-06

基于网络访问项序的移动用户重入网身份识别方法

王 征*，包 磊

（大连理工大学 软件学院，辽宁 大连 116024）

摘要： 用户身份的唯一性标识是任何移动商务营销活动所必不可少的一项基础工作.由于多种因素的交叉影响，移动用户大进大出已成为移动运营商面临的普遍现象.一旦用户重新入网，不仅针对旧号码的用户特征刻画记录将完全荒废，而且面向新号码的用户洞察又需要从头开始，并将耗费相当长的时间才能得到完整的用户画像.而另一方面，基于用户资料的移动用户身份识别准确率仅为42%.因此，针对移动用户重入网身份识别的问题，提出基于用户网络访问项序的用户相似性计算方法，通过数据预处理、相似用户集裁剪、用户身份识别等6个步骤来精确定位重入网用户身份.在某电信运营商某地区25 809个用户60 d网络访问日志这一数据集上，对所提方法进行了实验，总体准确率为98.32%，验证了方法的可行性与有效性.

关键词： 网络访问项序；移动用户；电信重入网；身份识别；用户相似性计算

中图分类号：TP391 **文献标识码：**A **doi：**10.7511/dllgxb201502016

0 引 言

移动互联网蕴藏着无限的商机，面向移动商务而进行的移动用户洞察与画像、超细分微营销、个性化商品推荐等活动可以使移动互联网的价值和魅力得以淋漓尽致地展现.而在所有这些移动商务营销活动背后，通过手机号来标识用户唯一性是一项必不可少的基础工作.然而，随着运营商新产品资费、营销活动、渠道酬金政策等因素的影响，移动用户选择弃卡重入网的情况广泛存在[1]，移动用户大进大出已成为移动运营商面临的普遍现象[2]；一旦用户重入网，不仅针对旧号码的用户特征刻画记录将完全荒废，而且面向新号码的用户洞察又需要从头开始，并将耗费相当长的时间才能得到完整的用户画像.因此，对重入网用户的身份进行有效识别，从而实现新旧用户身份的精确匹配，就成为移动互联网领域中亟待突破的一大关键问题.全球最大的移动运营商——中国移动通信集团公司将重入网用户的识别专门列为其经营分析系统中的一个重要的功能模块[3]，并提出需要建立精确的模型来识别.该问题的研究不仅有利于避免移动商务营销所面临的冷启动难题，而且对于公安机关有效识别犯罪分子的手机号码、对于个人隐私权边界的确定[4]，及对于移动运营商深入洞察用户重入网原因、制定相应的营销策略、降低用户重入网所导致的大量渠道运营成本及卡号资源浪费，都具有极其重要的现实意义.

纵观目前针对重入网用户身份识别问题的研究，代表性解决方案主要有：(1)通过用户资料来快速识别重入网用户.由于用户资料缺失、用户资料与用户本人不一致，以及完全不需要登记用户信息即可办理业务的预付费重入网等情况的存在，该方法识别用户身份的准确率仅为42%[5]，在实际操作中基本无法使用[6].(2)基于IMEI (international mobile equipment identity)码的识别技术.根据同一台手机在一段时间内使用了不同的手机号码，从而判定该用户为重入网用户.该方法对于大量更换终端而重入网的用户无能为力，而重入网用户的手机更换率高达32%以上[5]；同时，由于交换机或终端设备等原因，可利用的IMEI数据的准确性不高[7].(3)重入网用户

收稿日期：2014-08-05； **修回日期：**2015-01-08.
基金项目：国家自然科学基金资助项目(71271037,71272093,71201014).
作者简介：王 征*(1978-)，男，博士，副教授，E-mail:drwz@dlut.edu.cn.

的呼叫指纹识别方法.该方法的代表性研究成果包括美国麻省理工学院 Montjoye 等研究发现的“基于 4 个用户行为踪迹的参照要素,以相当粗糙的空间和时间辨识度,就足以识别和确认 95%用户的真实身份”[8],Calabrese 等建立的每个手机号码的移动轨迹模型[9],Mazhelis 等基于手机用户的环境和用户的手机操作行为的用户识别方法[10],何瑞江从用户呼叫指纹、IMEI 信息、用户资料等多个方面对重入网用户的身份识别进行的研究[6]等;借助呼叫指纹识别模型,可以有效解决由于用户资料缺失、用户更换手机等造成的重入网用户识别难题;然而,由于呼叫指纹识别方法基于用户语音详单及用户位置等数据而实现用户识别,该方法难以运用于目前大量存在的、几乎没有通话记录和位置记录的上网卡的识别.

通过研究问题特征发现,不同用户移动上网行为的差异化程度较高,每个用户所访问的网站分布、网站访问频次、网站访问时间顺序总会有一定的特异性,通过用户的网络访问日志来识别重入网用户身份是一条可行而有效的途径.因此,本文针对重入网用户身份识别问题,提出基于网络访问项序的移动用户重入网身份识别方法,并在某电信运营商的真实数据集上进行测试.

1 问题定义及其解决思路

重入网是指用户已经拥有某电信运营商的一个移动号码,由于某原因又买了该电信运营商的另一个移动号码入网,用新号码全部或者部分替代旧号码的现象.这样的用户称为重入网用户.重入网用户的识别需要对比新用户和其他所有用户在系统中记录的特征.该问题可进一步定义为:已知有用户库 $U=\{u_1,u_2,\cdots,u_m\}$,以及 U 中每个用户在一段时间 T 内的上网记录 $Record(u_i,T)$,$\forall u_i\in U$,当一个入网用户 v 以及他在短时间 T' 内的上网记录 $Record(v,T')$ 已知时,如何利用已有的用户库 U 及其上网记录来判断 v 与 U 中的某个用户 u_j 为相同身份的用户.若 $u_j=\varnothing$,则用户 v 为新入网用户.

结合对移动用户网络行为的研究,发现如下几条用户识别规则:

(1)不同用户访问的网站种类和网站数量往往不同;

(2)个性化强的网站通常只被很少的用户访问;

(3)每一用户在访问网站的时间上都呈现出一定规律;

(4)不同用户在访问多个网站的前后顺序上具有较高的差异化程度.

根据上述规则,建立了包含 6 个步骤的重入网用户身份识别方法(图 1),这些步骤处理过程的复杂度由浅入深,逐步缩小用户身份识别范围,一方面提高识别的准确度,另一方面将复杂度较高的算法步骤延迟到逐步缩小的数据集上执行,有利于提高算法的执行效率.由于该方法既考虑到了用户访问的网站(项目),又考虑到了访问的时间顺序,称该方法为“基于网络访问项序的重入网用户身份识别方法”.

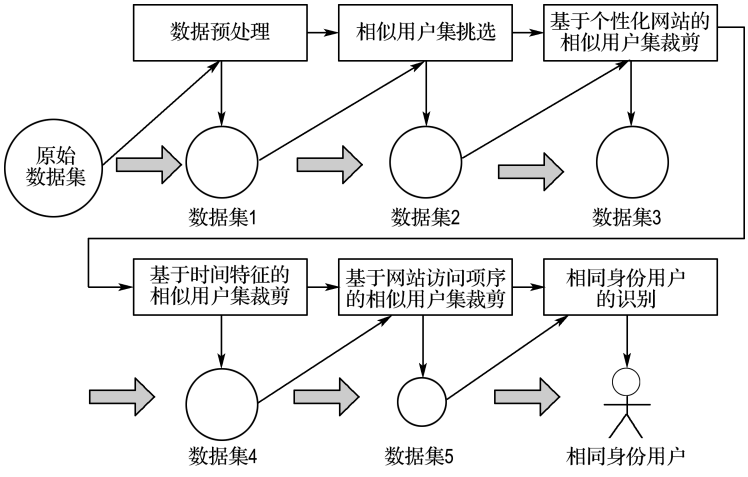

图 1 基于网络访问项序的重入网用户身份识别方法流程

Fig. 1 Procedure of rejoining user identification method based on the network access item and procedure

2　基于网络访问项序的重入网用户身份识别方法

(1)数据预处理

移动运营商的网络日志文件中记录着每个用户在某个时间点对某个 URL(uniform resource locator,统一资源定位器)请求的详细信息(50 余个属性).首先从庞大的原始日志中提取对问题研究有价值的字段——用户手机号、URL、访问时间.提取后的日志文件大大缩减(560 GB→110 GB),为后续计算工作减轻了负担.同时,按照如下的格式进一步提取每个用户在每天的网络访问项序:

{u, date, domains:{Website:{Website1, Website2, Website3, … }}, time:{time1, time2, time3,…}}

该记录表示用户 u 在日期 date 访问的网站 Website 有 Website1, Website2, Website3,…,访问的时间分别为 time1, time2, time3,….该种方式以 d 为单位统计出每个用户的上网行为,为后续计算中按 d 来比较用户网络访问行为奠定基础,而且可以很好地避免后续实验中一些数据的重复计算问题.

(2)相似用户集挑选

作为初次的"海选",由于涉及大量用户,相似用户集的挑选应在考虑方法执行效率的基础上尽量缩小相似用户的范围.为此,基于杰卡德相似性度量方法,使用下式作为两用户相似度的一个初步衡量:

$$SimiOne(v,u_i)=\frac{|D(v)\cap D(u_i)|}{|D(v)\cup D(u_i)|} \quad (1)$$

式中:$D(v)$ 和 $D(u_i)$ 分别代表新入网用户 v 和已知用户库 U 中用户 u_i 在一定时间内经常访问的网站集合.将用户经常访问的网站定义为网站访问天数占上网天数 80% 以上的网站.当 $SimiOne(v,u_i)$ 大于阈值 0.6(据实验而得)时,便将 u_i 加入 $SimiUsers(v)$ 这一与用户 v 相似的用户集中.

(3)基于个性化网站的相似用户集裁剪

由于式(1)仅是基于用户经常访问的网站个数比例进行的相似度计算,而在用户经常访问的网站中,有些网站是其他用户频繁访问的门户网站,而有些则是只有某几个用户访问的个性化网站,这种个性化的网站最能代表用户的网络访问行为特征,所以在相似度计算公式中应凸显这种个性化网站的作用.因此,将用户访问的个性化网站的权重考虑进来,使用下式来进一步计算两个用户的相似度:

$$SimiTwo(v,u_i)=\sum\frac{1}{V(d)};\ d\in D(v)\cap D(u_i) \quad (2)$$

其中 $V(d)$ 为访问过域名 d 的用户总数.

式(2)中,由于访问用户数越少的网站,个性化程度越高,其权重也应越大,将网站的权重公式定义为所有用户访问数的倒数.由式(2)可得到用户 v 和 $SimiUsers(v)$ 中每个用户的新的相似度,根据这一相似度计算结果倒序排序,取一定比例(前 5%,据实验而得)的用户作为裁剪后的 $SimiUsers(v)$ 用户集.

(4)基于时间特征的相似用户集裁剪

在每一用户的整体上网时间方面,用一个 24 维的向量 $\boldsymbol{h}=(h_1\quad h_2\quad\cdots\quad h_{24})$ 表示用户的上网时间特征,其中 h_i 表示第 $(i-1)$ h 到第 i h 该用户的网络访问天数.如果在第 $(i-1)$ h 到第 i h 内至少有 20 min 有 URL 请求的发送,则给 h_i 累加 1.以上网时间相似度 0.69(据实验而得)为标准,采用余弦相似度计算方法,将新用户 v 的相似用户集 $SimiUsers(v)$ 中所有和 v 的上网时间相似度 $\leqslant 0.69$ 的用户全部过滤掉.

同时,针对用户经常访问的网站计算出用户单独针对这些网站的访问时间特征,基于这一特征对相似用户集进行进一步裁剪.对于每个经常访问的网站 w 同样用一个 24 维的向量 $\boldsymbol{h'}=(h'_1\quad h'_2\quad\cdots\quad h'_{24})$ 表示该用户对网站 w 的访问时间特征,其中 h'_i 同样表示第 $(i-1)$ h 到第 i h 网站 w 被用户访问的天数.以 0.53(据实验而得)为标准,采用余弦相似度计算方法,将新用户 v 的相似用户集 $SimiUsers(v)$ 中所有和 v 的常访问网站的访问时间相似度 $\leqslant 0.53$ 的用户全部过滤掉.

上述两个过滤过程放入 $SimiThree$ 函数执行.

(5) 基于网站访问项序的相似用户集裁剪

用户网络访问项序反映了用户兴趣的动态变化,可以用来作为识别用户身份的一个重要标志.如用户 u_1、u_2、u_3 各自访问网站的先后顺序排列如下:

$$u_1:w_1\rightarrow w_4\rightarrow w_5\rightarrow w_3\rightarrow w_6\rightarrow w_2$$

$$u_2:w_2\rightarrow w_5\rightarrow w_4\rightarrow w_6\rightarrow w_5$$

$$u_3:w_4\rightarrow w_5\rightarrow w_3\rightarrow w_6\rightarrow w_2$$

从以上访问内容看，用户 u_1、u_2、u_3 都访问了网站集合$\{w_2, w_4, w_5, w_6\}$，但从访问项序来看，u_3 和 u_1 的子序集“$w_4 \rightarrow w_5 \rightarrow w_3 \rightarrow w_6 \rightarrow w_2$”是一致的，而 u_2 和 u_1 的访问顺序大相径庭，没有任何一个子序集相同，因此用户 u_1 和 u_3 更为相似.

更进一步，对两个用户 u_i 和 u_j 在网络访问项序上的相似性进行如下定义：设用户 u_i 和 u_j 在一段时间内的网络访问项序分别为

$$u_i: w_{i1} \rightarrow w_{i2} \rightarrow \cdots \rightarrow w_{im} \rightarrow \cdots \rightarrow w_{in} \rightarrow \cdots \rightarrow w_{ip}$$

$$u_j: w_{j1} \rightarrow w_{j2} \rightarrow \cdots \rightarrow w_{jm} \rightarrow \cdots \rightarrow w_{jn} \rightarrow \cdots \rightarrow w_{jq}$$

若上述两个项序中的子序集 $w_{im} \rightarrow \cdots \rightarrow w_{in}$ 和 $w_{jm} \rightarrow \cdots \rightarrow w_{jn}$ 完全相同，则集合$\{w_{im}, \cdots, w_{in}\}$中元素的个数就是这一子序集相似性的一个度量；当匹配到相似子序集之后，该子序集就从两个用户的访问项序中删除，并继续寻找其他相似的子序集；如此迭代，直至没有相似子序集存在，则所有相似子序集度量之和就是两个用户网络访问项序的相似性.

以 14.3%（据实验而得）为阈值，将相似用户集 $SimiUsers(v)$ 中所有和 v 的网络访问项序相似度排名 $<14.3\%$ 的用户全部过滤掉. 将这一裁剪过程放入 $SimiFour$ 函数执行.

(6) 相同身份用户的识别

通过上面几步，若 $|SimiUsers(v)|=0$，则判断用户 v 为新入网用户；若 $|SimiUsers(v)|=1$，则直接将该用户作为与用户 v 具有相同身份的用户而返回；若 $|SimiUsers(v)|>1$，则需要进一步从用户集中挑选与用户 v 最相似的用户作为相同身份的用户.

由于 $SimiOne$ 与 $SimiTwo$ 都是在用户访问网站相似性方面的衡量，使用式(3) 对这两个相似度进行综合计算，得到 $SimiOneTwo$. 用户访问的网站、时间特征、访问项序等方面对于用户相似度计算的贡献很难分出伯仲，因此通过平均法计算出每一用户与入网用户 v 的总相似度 S. 将 $SimiUsers(v)$ 中的所有用户按照 S 值降序排序后，选取前 $N(N \geqslant 1)$ 个作为与入网用户 v 具有相同身份的用户，具体计算规则为：所有用户按照 S 值降序排序，除排名第一位的用户外，其他任何用户 u 的 S_u 值若大于 $S_{u-1} \times (1-5\%)$，且用户 $u-1$ 也被作为相同身份的用户而输出，则将用户 u 加入结果集 $SameUsers(v)$ 输出. 若 $SameUsers(v)$ 集合中的用户多于 1 个，则可通过查看用户资料等其他办法辅助确定用户身份.

$$SimiOneTwo(v,u_i)=\frac{\sum W(d)}{\sum W(e)};$$

$$d \in D(v) \cap D(u_i),\ e \in D(v) \cup D(u_i) \quad (3)$$

上述算法的 5% 根据经验而得，因为存在多种情况可能使 $SimiUsers(v)$ 中与 v 具有相同身份的用户并不能排在首位，而上述方法能在一定程度上保证仍能够将该用户返回.

3 实验结果及其分析

为验证基于网络访问项序的重入网用户身份识别方法的可行性与有效性，针对某电信运营商提供的某地区 25 809 个去隐私用户从 2013-11-04 到 2014-01-02 时间段内(60 d)的网络访问日志，使用所提方法进行了实验，实验环境、思路及结果如下：

(1)实验环境及思路

本文提出的用户身份识别方法的所有细节都被编码为 Java 程序，整个测试过程运行在操作系统为 Linux，配置为 Pentium Ⅳ 2.4 GHz，8 GB 内存的台式计算机上.

将所有用户在 60 d 里的数据集按照图 2 的思路进行了拆分，针对 25 809 个用户后 t d 的网络访问数据，在 20 809 个用户的前$(60-t)$ d 中查找相同身份的用户. 由于有 5 000 个用户在用户库 U 中不存在相同身份的用户，而另外 20 809 个用户在用户库中存在，使用如下 3 个表达式来判断身份识别方法的准确率：

$$\frac{\text{预测 5 000 用户不在库中的用户数}}{5\,000}\times 100\%$$

$$\frac{\text{预测 20 809 用户在库中的用户数}}{20\,809}\times 100\%$$

$$\frac{\text{预测 5 000 用户不在库中的用户数}+\text{预测 20 809 用户在库中的用户数}}{25\,809}\times 100\%$$

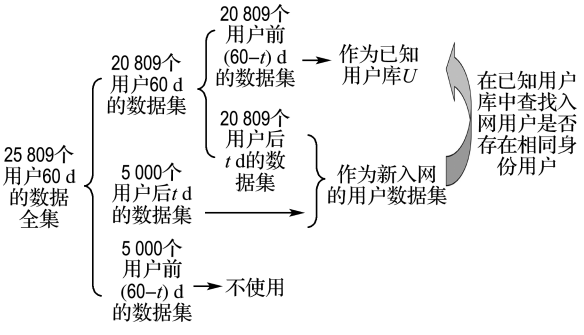

图 2 数据集的拆分及实验思路

Fig. 2 Experiment method of dataset split

由图 2 知，应该给出一个最小的 t 值，使方法仍能够以较高的准确性识别用户. 关于 t 值，进行了初步实验，并发现，随着 t 的增长，识别方法的准确率也在逐渐增长，当 t 从 4 增长到 5 时，识别方法的准确率有一个大幅增长，但当 $t>5$ 之后，识别方法的准确率涨幅不大，因此选择用户 5 d 的网络访问数据作为重入网用户身份识别的源数据.

(2)实验结果

针对一个入网用户 v，在 20 809 个用户的数据集上逐步骤地运行了“基于网络访问项序的重入网用户身份识别方法”. $SimiUsers(v)$ 用户集在本文方法每一步骤运行之后的变化趋势(根据式 $\frac{|SimiUsers(v)|}{20\ 809}$ 计算而得)及运行的时间如表 1 所示. 由表可见，本文方法的第 2 步骤实现了对相似用户集的大幅度裁剪；但由于用户数的下降、用户间相似性的提高，第 3～5 步骤的裁剪幅度有所下降，这些步骤的算法复杂度有所提高，其主要工作更侧重于用户身份识别的精度.

当算法完全执行完毕后，根据前述 3 个表达式得到了如表 2 所示的 3 个准确率，它们都在 98%以上. 而算法在经过第 1 步数据预处理之后，针对一个用户识别的总体运行时间仅为 6.127 s，算法的准确率与效率得到了较好的平衡，并可以被付诸于工程实践应用当中.

表 1　用户身份识别方法各步骤 $SimiUsers(v)$ 用户集数量的变化以及运行时间

Tab. 1　Variation of user number in the set $SimiUsers(v)$ and running time of each stage of the user identification method

步骤	裁剪后比例/%	运行时间/s
2	14.120 000	4.900
3	0.007 000	1.100
4	0.003 000	0.067
5	0.000 429	0.032

表 2　基于网络访问项序的用户身份识别方法的准确率

Tab. 2　Accuracy rate of the user identification method based on the network access item and procedure

预测不存在用户的准确率/%	预测存在用户的准确率/%	预测所有用户的准确率/%
99.06	98.14	98.32

4　结　语

移动用户重入网之后的身份识别问题是各项移动商务营销活动有效开展的基础问题. 针对这一问题，本文提出了基于网络访问项序的重入网用户身份识别方法，从用户网络访问的网站特征、时间特征、项序特征这几大方面，逐步精化重入网用户的身份识别手段，从而实现对重入网用户身份精确定位的最终目标. 该方法被运用于某电信运营商某地区 25 809 个用户 60 d 的网络访问日志数据上，实验的总体准确率为 98.32%.

本文提出的基于网络访问项序的重入网用户身份识别方法实现了用户身份识别效率与准确率的较好平衡，不仅为移动用户重入网之后的身份识别提供了新思路，而且对于解决移动商务营销所面临的冷启动难题，对于公安机关有效识别犯罪分子的手机号码，对于个人隐私权边界的确定，对于移动运营商深入洞察用户重入网原因、制定相应的营销策略、降低用户重入网所导致的大量渠道运营成本及卡号资源浪费，都具有重要的现实意义.

参考文献：

[1]　罗　亚. 移动电话用户重入网识别及营销建议[D]. 北京：北京邮电大学，2010.
LUO Ya. Identification of the re-joined mobile customer and marketing proposals [D]. Beijing: Beijing University of Posts and Telecommunications, 2010. (in Chinese)

[2]　张大亮. 移动用户新增来源与流失去向分析及其应用探究[D]. 呼和浩特：内蒙古大学，2012.
ZHANG Da-liang. The analysis and application research for mobile user′s adding source and lossing whereabouts [D]. Hohhot: Inner Mongolia University, 2012. (in Chinese)

[3]　中国移动通信集团公司. 中国移动省级 NG2-BASS(v4.5)业务规范[S]. 北京：中国移动通信集团公司，2013：147-148.
China Mobile Communications Corporation. Business Specification for New Generation Business Analysis Support System [S]. Beijing: China Mobile Communications Corporation, 2013: 147-148. (in Chinese)

[4]　朱　慧，刘洪伟，陈　丽，等. 网络用户的信息隐私边界及其敏感度等级研究[J]. 广东工业大学学报，2013，**30**(4)：26-32.

ZHU Hui, LIU Hong-wei, CHEN Li, *et al*. Research on information privacy boundaries and sensitivity of net users [J]. **Journal of Guangdong University of Technology**, 2013, **30**(4):26-32. (in Chinese)

[5] 艾 达,罗爱平. 移动通信重入网用户识别算法分析研究[J]. 西安邮电学院学报, 2012, **17**(3):30-33.
AI Da, LUO Ai-ping. Study on mobile communications rejoin customer identification algorithms [J]. **Journal of Xi'an University of Posts and Telecommunications**, 2012, **17**(3):30-33. (in Chinese)

[6] 何瑞江. 利用呼叫指纹挖掘电信重入网客户[D]. 兰州:兰州大学, 2009.
HE Rui-jiang. Using fingerprint recognition technology mining telecommunication reentry net customer [D]. Lanzhou: Lanzhou University, 2009. (in Chinese)

[7] 章昱梓. 移动用户重入网分析系统的分析与设计[D]. 北京:北京邮电大学, 2011.
ZHANG Yu-zi. The analysis and design of the mobile users rejoin network analysis system [D]. Beijing: Beijing University of Posts and Telecommunications, 2011. (in Chinese)

[8] de Montjoye Y A, Hidalgo C A, Verleysen M, *et al*. Unique in the Crowd: The privacy bounds of human mobility [J]. **Scientific Reports**, 2013, **3**: 1376.

[9] Calabrese F, Diao M, Di Lorenzo G, *et al*. Understanding individual mobility patterns from urban sensing data: A mobile phone trace example [J]. **Transportation Research Part C: Emerging Technologies**, 2013, **26**:301-313.

[10] Mazhelis O, Puuronen S. A framework for behavior-based detection of user substitution in a mobile context [J]. **Computers and Security**, 2007, **26**(2):154-176.

An identification method of rejoining mobile user based on network access item and procedure

WANG Zheng*, BAO Lei

(School of Software Technology, Dalian University of Technology, Dalian 116024, China)

Abstract: User uniqueness identification is a kind of necessary work for any marketing activity in mobile commerce. However, as a result of many factors and their interactions, it is a common phenomenon for mobile operators that a great amount of mobile users enter communication network and exit. Once users re-enter the communication network, not only the descriptions of old users are totally useless, but also the work of inspecting new users has to be re-started and the whole user profile can only be obtained after a very long time. On the other hand, the accuracy rate of the user identification work based on user information is only 42%. Therefore, the mobile user identification problem is focused after they re-enter the communication network. And a similarity calculation method is presented based on users' network access item and procedure. The method can identify a user by 6 steps from pre-processing data → trimming the similar user set → identifying users, etc.. Experiments are made on a real data set of the network access log of 25 809 users from a communications corporation in 60 days. The overall accuracy rate of the presented identification method is 98.32%, which shows the feasibility and effectiveness of the method.

Key words: network access item and procedure; mobile user; rejoining communication network; identity recognition; user similarity calculation